PROGRESS IN STEREO MAPPING

H. Harlyn Baker, Thomas O. Binford, Jitendra Malik, Jean-Frederic Meller
Artificial Intelligence Laboratory, Computer Science Department,
Stanford University, Stanford 94305.

1: Introduction

Rather than a description of a single piece of research, this is more in the line of a collective report on some areas we've been addressing in our research in stereo mapping. We have been developing tools and experimenting with matching strategies that will build to a rule-based stereo matching system. In particular, we have been:

a) demonstrating the design and the utility of the rule-based approach to surface inference from monocular information through the hand synthesis of matching strategies;

b) developing tools to support a mapping interactive test facility;

c) experimenting with the [Baker 1981] stereo mapping system, preparing to run it on some new imagery.

In a), we have been carrying out research aimed at the analysis and synthesis of rules for inference of three-dimensional shape from single images. We have been addressing inference of matching rules and the use of model-based analysis both with theoretical analyses and with hand and automated analyses of specific matching strategies; the latter applied to both real and synthesized imagery examples. This inference also has application in constraining search for matches in stereo correspondence.

Our work in developing tools has centered on:

a) a system for the hand construction of edge descriptions from hard copy imagery;

b) an interactive system for determining the transform to bring image pairs into collinear epipolar registration.

Both of these tools make extensive use of interactive graphics, and the latter takes much advantage of previous stereo research from our laboratory ([Gennery 1980]).

In c), we have been undertaking to apply the [Baker 1981] system to some new imagery. In this we hope to demonstrate its effectiveness, to expose its limitations, and to suggest both its role in an advanced mapping system and complementary research needed to improve its utility. Significant restructuring of the system was called for in enabling it to process this new imagery. Details of these changes are described in section 4, which deals with the matching process. Modifications have now been implemented, enabling the system to:

- function on the output of an improved edge operator [Marimont 1982];
- use *edge extent* as one of its parameters in seeking optimized correspondence;
- exploit prepared transform information in processing images whose epipolar lines are not collinear with the scanning axes of the cameras.

We will describe results in these areas of the research through discussion of the following:

a) the use of image edge descriptions produced using the digitizing facility and from an automated process [Marimont 1982] in synthesizing rules for stereo matching;

b) development of a system for epipolar registration of image pairs;

c) modifications to the [Baker 1981] stereo system (results later).

2: Inference and Modelling

2.1 Preamble – the digitizing facility

Our approach to rule development begins with hand synthesized and some automatically generated edge data. We have systems for the automatic generation of edge data (*i.e.* [Marimont 1982]). This data, extracted from a sufficiently wide selection of imagery types, gives good insight into the current capabilities of automated processes. Automated processes, however, are not able presently to give as meaningful a description of an image as we would like, and have not been designed to provide the aggregated abstractions research systems ([Lowe 1982]) will be soon supplying. To bridge this inadequacy, we work with both automatically generated data (the current state-of-the-art), and hand generated data (representative of the next generation of edge analysis processes). The hand generated data is obtained from a manually operated digitizing tablet. We have written a graphics-based digitizing and editing system to run with a GTCO tablet in producing these image descriptions.

Figure 2-1 below shows an image pair of a building complex (referred to as the Sacramento imagery). Figure 2-2 shows the results of tablet edge extraction on these images. Figure 2-3 shows the results of the Marimont operator [Marimont 1982] on the image pair. Manually generated edge data was produced using this facility for the analysis of rule synthesis of section 2. It was also used to digitize the building data of figures 2-1 for input to the OTV inference process, as figure 2-5.

2.2 Data for Rule Synthesis

We have obtained extended edge data from hand and automated processing for use in synthesis of matching rules. Results from earlier work on OTV analysis (orthogonal trihedral vertices) have been exploited [Perkins 1968] for rule formation in shape inference and in constraining search for correspondence. We have taken examples from the modelling of generic structures to produce ground and aerial views of a building complex, and have used this, as well as other data, in rule synthesis.

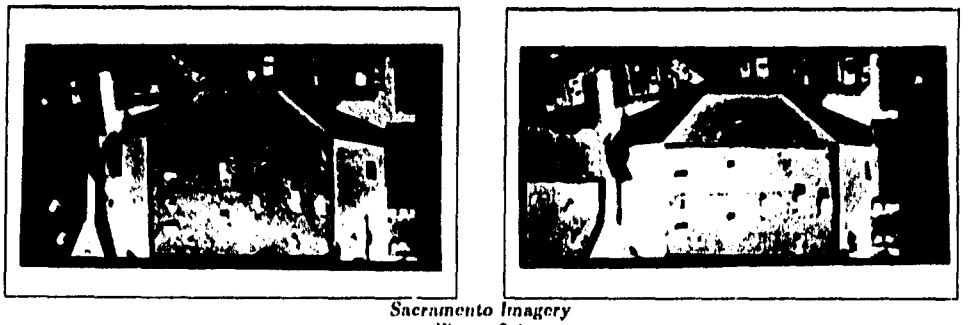
2.3 Modelling and Vision

2.3.1 – Modelling, prediction and interpretation

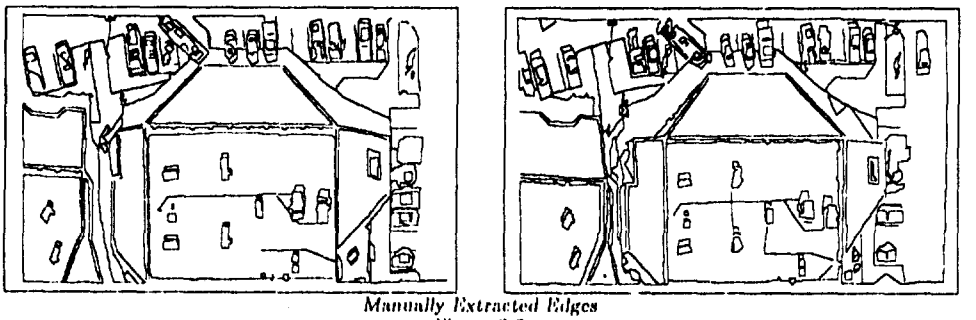
Of course, one of the primary goals of research in computer vision is the development of systems that can recognize and locate objects in images. In order to identify such an object, it is clearly necessary to have some description of its characteristics that can be detected in an image. A *representation* of an object is the form this description takes.

One approach to representation is to provide the system with three-dimensional models of objects. Rotation of these models will allow objects to be observed, conceptually, from differing viewpoints. If parameters in a particular model are allowed to vary it is possible to have that single model represent a whole class of objects; constraining the parameters functions to delimit sub-classes. Further model manipulations, such as partitioning and projection, can be used to aid in mapping model to imagery data. The information contained in such object models may be used to determine possible interpretations of image features (*e.g.* edges, ribbons, corners) and to provide feedback to predict the locations of such features in an image.

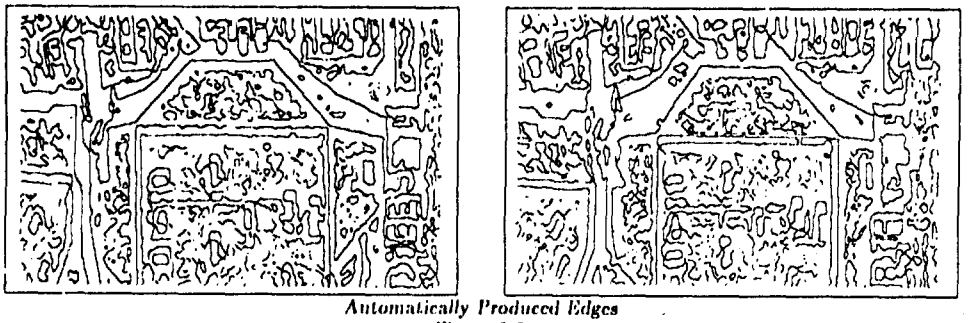
ACRONYM [Brooks 1981] is a three-dimensional rule-based modelling/vision system developed here at Stanford that provides, among other things, such feature prediction, model manipulation, and image inter-



Sacramento Imagery
Figure 2-1



Manually Extracted Edges
Figure 2-2



Automatically Produced Edges
Figure 2-3

pretation. The rule-base operates on the models and on the sensed data to accomplish scene interpretation. Such a rule-based approach has been shown to be an effective form for constraint and search implementation, and allows easy modification and addition of new rules without the need of altering the underlying code.

Our group's intention over the next few years is to build a rule-based stereo system operating within ACRONYM whose functioning will include model-based prediction. Working toward this, we have been carrying out experiments on scene inference and model-based prediction that will lead to a repertoire of stereo matching rules.

2.3.2 – Models and stereo matching

One of the major difficulties in determining stereo correspondence is in dealing with the large number of matches that are possible. Solution is generally found by search through a large parameter space, where possible correspondences are limited by geometric or photometric constraints. Search can be reduced even more dramatically by endowing the matcher with broad domain specific and domain independent knowledge. Such knowledge can be rule-based and model-based. Our proposition here is that the three-dimensional information in object models, along with inference and prediction mechanisms, can be used to interpret features in image pairs. These interpretations can then be used as filters to constrain the matching. We demonstrate this notion with the example of Orthogonal Trihedral Vertices, often referred to as cube corners or OTVs. Other rules synthesized from analysis of both manually extracted and automated edge processes follow.

The work on cube corners points to additional usefulness for a model-based approach. OTV orientation analysis (from matches across pairs of views) yields almost complete solution for camera parmameters; constraints on sizes (again, from rules and models) could complete the camera solution. But the orientation information yielded by a match of a pair of vertices is valid only if the vertex is a cube corner; thus it is necessary to be able to distinguish between vertices that are cube corners and those that are not. If the models contain sufficient information to identify cube corners, then the problem of determining cube corners independently of the identification process is eliminated. In fact, both the search for cube corners and the search for identification are likely to be reduced when they are combined.

2.3.3 – OTV rule-based analysis

OTV theory

In cultural scenes, we find a large number of interior and exterior corners of cubes – typically when two walls at right angles meet the roof or the floor. The importance of utilizing this common structural element – the Orthogonal Trihedral Vertex (OTV) – has been emphasized earlier[Lieber 1981]. Since they provide a very tight constraint – the three edges are mutually orthogonal in space – it is possible to calculate the three dimensional orientations given the projections in the image. This can be done for both orthographic and perspective viewing.

If the eye is assumed to be focussed on the vertex of the cube corner, perspective can be ignored and the projection of a cube corner in XYZ space will simply be its orthogonal projection on the XY plane.

Suppose that some 3 - star has angles between its rays a, b and c and also that the rays are represented by the unit vectors v_1, v_2, v_3. We are interested in detecting whether there are three vectors in XYZ space, which are mutually orthogonal and project, respectively to v_1, v_2, v_3.

Since projection is accomplished by dropping the z component, any 3 such vectors must be of the form $v_1 + \lambda_1 z$, $v_2 + \lambda_2 z$, and $v_3 + \lambda_3 z$ where z is the unit vector in the z direction.

Requiring mutual orthogonality implies that the dot products of these vectors in pairs be zero. From these conditions and some simple manipulations we can calculate the formulas for

$$\lambda_1 = \pm\sqrt{-\frac{(\cos a)(\cos c)}{(\cos b)}}, \quad \lambda_2 = \pm\sqrt{-\frac{(\cos a)(\cos b)}{(\cos c)}}, \quad \lambda_3 = \pm\sqrt{-\frac{(\cos c)(\cos b)}{(\cos a)}}$$

Hence solutions exist if

a) $\cos a$, $\cos b$, $\cos c$ are all non-zero and

b) either one or three of $\cos a$, $\cos b$, $\cos c$ are negative, so that the quantities under the square root sign are positive.

These results were first derived in [Perkins 1968].

Thus we have a way of both eliminating false candidates for being OTVs and finding the 3-D orientations of valid OTV's. This algorithm has been implemented and run on data from the digitizing tablet.

OTV with/from models

Our analysis begins on both images, processing bottom up on the two images separately. As the rule system identifies likely OTVs in images (from its models), it proceeds to match them. The system should already have a tentative identification of the buildings containing the OTVs, so there should be relatively few possible matches at this point. Only OTVs that could be the same point on the same object need be compared. The analysis results in depths of matched objects, for all those objects having OTVs.

This requires that the modelling system handle point elements, and that it include both:

- inferring OTVs from models (volumes);
- accessing OTVs stored explicitly with the models.

2.4 Rule Synthesis

2.4.1 – Inference rules

We continue with the development of inference rules. This work is a logical extension of previous work [Binford 1981, Lowe 1982] done at Stanford in developing rules for inferring surface information from a single view. General assumptions about illumination, object geometry, the imaging process etc. have been used to derive rules for making specific inferences. For stereo vision Arnold and Binford [Arnold 1980] have developed conditions on correspondence of edge and surface intervals. We divide our rules into two categories: monocular rules, which enable surface inference from a single view; and stereo rules, which facilitate cross-image matching.

2.4.2 - Monocular and stereo rules.

1. Monocular rules – Rules which have been developed for inferences from monocular views can be utilised to provide a partial 3-dimensional interpretation which directs search in the second view. This category includes the rule for interpretation of Orthogonal Trihedral Vertices.

   Another example is the T-junction rule [Binford 1981] which states that *'In absence of evidence to the contrary, the stem of a T is not nearer than the top, i.e. is coincident in space or further away'*. Application of this rule gives a set of nearer/farther relations. A hypothesized correspondence of edges which leads to inconsistent conclusions from the two views can be pruned from the search.

   An image line which is straight must be the image of a straight space curve unless the curve is planar and the observer is coincidentally aligned with the plane of curvature. This enables us to dismiss correspondences between straight edges in one view and curves in the other view. If two image curves are projectively consistent with parallel, we assume they are images of curves which are parallel in space. That implies that their images in the other view would be parallel *i.e.* parallels map to parallels.

   As these examples illustrate, most of the rules in [Binford 1981, Lowe 1982] and others developed by Malik and Binford have as direct corollaries stereo rules for checking the legality of a match. They can even direct the search process.

2. Stereo Rules - these are rules which have been derived from the stereo imaging process, and are a function of the imaging geometry.

   An example rule in this class, which has long been used for finding stereo correspondences, is the epipolars rule – corresponding points must lie on corresponding epipolar lines. These rules have inherently no monocular analogs. Here are a few:

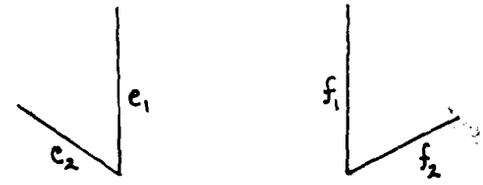
   a) Horizontal planes in one view get mapped to horizontal planes in the other view.

   b) Use of projective and quasi-projective invariants. This has not been examined in detail. Duda and Hart[Duda 1973] devote a chapter to this topic which has not really been exploited in stereo work.

   c) Conditions on correspondence of edges and surface intervals[Arnold 1980].

   d) Surface Occlusion rules:

      Surfaces visible in one view can be occluded in the other view. We are interested in the conditions when this takes place. The basic idea is that if we cross a surface, an obscuration of edge occurs. A left surface visible in a right view is visible in the left view unless there is obscuration by a tall object. Similarly a right surface visible in a left view will be seen unless obscured by a tall object. These surface-obscuration rules can be formalized by the cross-product rule:



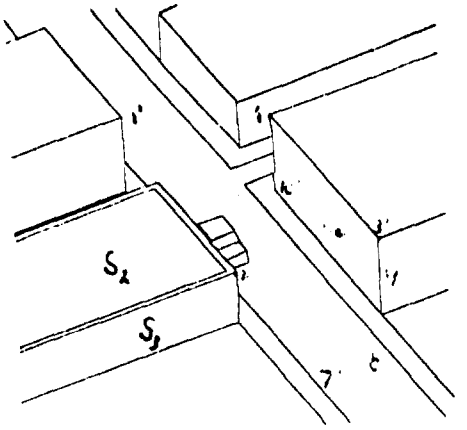
      For the hypothesized edge match e_1 with f_1 and e_2 with f_2, we compute the Z-component of the vector cross-product in the left image pair and the right image pair. If the z-components have opposite signs, we are seeing opposite sides of the surface. That implies that the object is not opaque.

2.4.3 – Use of inference rules in a test analysis

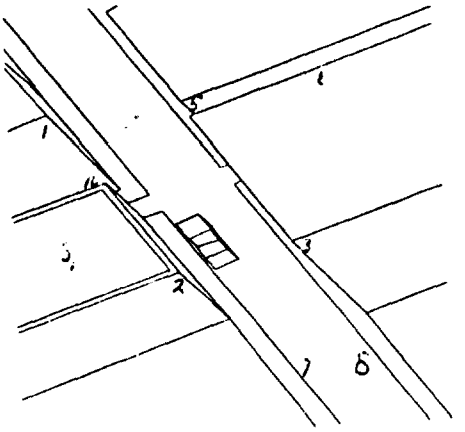
Our preliminary results indicate good potential for the success of this approach. On hand simulations with line drawings of stereo pairs, the rules helped narrow down the choices considerably.

Consider the imagery shown in figures 2-4 and 2-5. Figure 2-5 is the right view and 2-4 the left view. Vertices 1, 2, 3 are orthogonal trihedral vertices. Using the formulae developed earlier, we can find the 3-D orientations of the edge vectors. These can be matched with the 3-D orientations of 1', 2', 3' to obtain a registering of these vertices when combined with the epipolar constraint. All OTV's in one view need not be visible in the other i.e. 4'. Of the monocular constraints, the other major constraints which can be seen here are the T-junction rule and

the parallels rule. In figure 2.5 edge 5 is behind edge 6. Edges 7 and 8 are parallel and so are 7' and 8'. A match of 8 with 9' would not be accepted. Surfaces S_1 and S_2 are both horizontal planes (as can be deduced from the OTV analysis) and can be matched. Surface S_3 is not horizontal. Here of course, this does not provide any new information. Surface S_4 being a left face in a left view is not guaranteed to be visible in the other view as in fact it is. The cross-product rule could be used to dismiss a match between 10 and 10'.



Description of Left Image of Stereo Pair
Figure 2-4



Description of Right Image of Stereo Pair
Figure 2-5

3: Image Registration

3.1 Introduction to Epipolar Geometry

The search process in automated stereo mapping can be greatly restricted, and computation times significantly reduced, if information is available relating the relative camera geometries of a stereo pair. Often this information is available in the reconnaissance data (or at least a rough approximation to it). Other manual and automated schemes have been devised to provide the information when it is not present with the imagery (see [Hallert 1960], [Gennery 1980]). This camera geometry information allows establishing epipolar correspondence of lines across images. When this has been done, search in one image for match points of a feature in the other image can be constrained along a single vector. More generally, any features lying along a particular vector in the one image may be found along a single vector in the other image. These image plane vectors are termed epipolar lines. Corresponding vectors are termed corresponding or conjugate epipolar lines.

In this section we detail an algorithm for determining conjugate epipolar lines in a set of imagery for which such camera geometry information is not explicitly available. Here, we rely upon an operator to select corresponding points in the two images. The system automatically improves the resolution of the correspondence through Fourier interpolation over a match window [Gennery 1980]. The set of such points is taken by an automated camera solver to produce the needed geometric information. This point selection is done with pan/zoom cursor control on a graphics device. If the camera information is available (as, for example, from reconnaissance data), then the point matching phase may be omitted (although this provision has not been enabled in the current system). Equally, rough camera geometry information, if available, may be used to partially automate the point selection phase, although again this is not implemented here. [Gennery 1980] and [Moravec 1980] have implemented totally automatic camera solvers in their stereo matching systems. Our next improvement to this registration system will be to incorporate the image sampling and feature matching of the [Gennery 1980] system, removing the need for manual point selection.

3.2 Glossary of Terms

Given two cameras C_1 and C_2 with origins O_1 and O_2 and focal planes P_1 and P_2, we call:

Bundle of lines: A family of lines containing a common point.

Epipolar coordinates of a point: The number of the epipolar line it belongs to, and its distance to a fixed reference, like the epipole if it exists.

Epipolar direction: The direction of all epipolar lines if they are parallel.

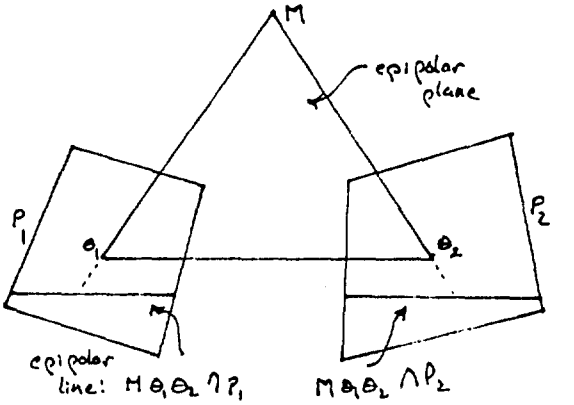
Epipolar line: The intersection of an epipolar plane with a focal plane. Alternate definition: the image in one camera of the pre-image of a point in the other camera's focal plane.

Epipolar plane: Any plane containing the two camera centers O_1 and O_2.

Epipolar space: A space where the coordinates are the epipolar coordinates. In this space, a horizontal line is an epipolar line, and the epipole, if it exists, is a whole vertical line.

Epipole: The intersection, if it exists, of all epipolar lines in a focal plane.

Conjugate epipolar lines: the intersections of an epipolar plane with the two focal planes.



Epipolar Geometry
Figure 3-1

3.3 Background Theory

Given two stereo images P_1 and P_2 (the content of the focal planes P_1 and P_2 of the cameras), and two lines L_1 and L_2 contained in P_1 and P_2, in general every point of L_1 maps to a line segment in P_2, and

there is no particular relationship between the line segments mapping to different points.

Fundamental property:

Every point in L_1 will map to a line segment contained in the same line L_2 if, and only if L_1 and L_2 are a pair of conjugate epipolar lines.

Epipolar Geometry:

The family of epipolar lines in a focal plane is:

Either

a) a set of parallel lines having a common epipolar direction,

or

b) a bundle of lines, the intersection of which is the epipole.

3.4 Requirements of the System

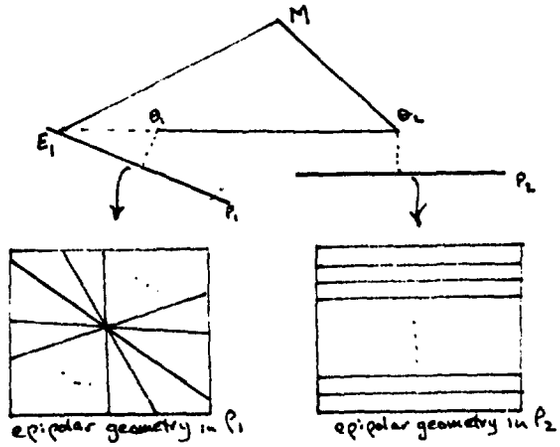
Given a pair of stereo images, we want to:

1) identify the kind of epipolar geometry present in the images;

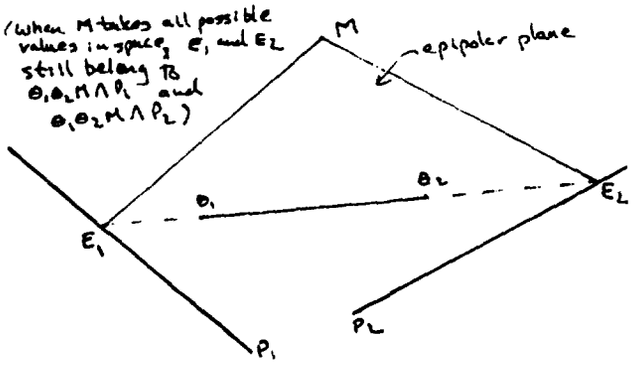
2) explicitly show the epipolar lines belonging to each image;

3) for each image, compute the parameters which relate the original coordinates to the epipolar coordinates;

4) construct the image transforms in epipolar space.



Top view of a situation with only one epipole
Figure 3-2



Top view of a situation with epipoles
Figure 3-3

Prior to these 4 steps, we will need to solve for the cameras, that is, to determine the 5 parameters describing their relative orientation. A procedure developed at Stanford [Gennery 1980] is used for this.

3.5 Algorithm Used

3.5.1 – Camera registration output

Each camera is viewed as a referential (O_i, x, y, z), $i \in \{1,2\}$. The registration procedure yields *azimuth, elevation, pan, tilt,* and *roll* of one camera with respect to the other. From there, we compute:

1) the rotation matrix R between the two referentials:

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}$$

2) the translation unit vector t, the components of which are:

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \text{in the base } O_1xyz, \quad \begin{pmatrix} \lambda \\ \mu \\ \nu \end{pmatrix} \text{in the base } O_2xyz$$

Note that the magnitude of the translation vector cannot be determined from a pair of images.

3.5.2 – Epipolar geometry determination

The focal planes P_i are planes parallel to O_ixy, intersecting O_iz at $z = f_i$, the focal distance. There is an epipole in plane P_i if and only if the translation vector intersects this plane, that is, if and only if its third component is not zero.

We thus determine the case we are working with:

if $\gamma \neq 0$ and $\nu \neq 0$, there are two epipoles:	CASE 1
if $\gamma \neq 0$ and $\nu = 0$, there is one epipole E_1:	CASE 2
if $\gamma = 0$ and $\nu \neq 0$, there is one epipole E_2:	CASE 3
if $\gamma = 0$ and $\nu = 0$, there are no epipoles:	CASE 4

Remarks:

a) $\gamma = 0$ is replaced in the code by $|\gamma| <$ threshold, where threshold is chosen as a function of the arithmetic precision of the machine: if we had infinite precision, then we could consider every case as being case 1. Here threshold = .0001 was found to be a good estimate.

b) Most image pairs will belong to the first case, with γ and ν of the order of .1. The epipoles exist, are outside the picture frame, and, for the images worked with to date, tend to be at a distance of about 10 times the picture dimension.

3.5.3 – Epipoles and epipolar directions

If the epipole E_i exists, it is the extremity of the vector collinear to the translation vector, with a third component equal to f_i. If it does not, then the translation vector is the epipolar direction. Hence:

Case 1:	$E_1(\frac{\alpha f_1}{\gamma}, \frac{\beta f_1}{\gamma})$	$E_2(\frac{\lambda f_1}{\nu}, \frac{\mu f_1}{\nu})$
Case 2:	$E_1(\frac{\alpha f_1}{\gamma}, \frac{\beta f_1}{\gamma})$	$V_2(\lambda, \mu)$
Case 3:	$V_1(\alpha, \beta)$	$E_2(\frac{\lambda f_2}{\mu}, \frac{\mu f_2}{\mu})$
Case 4:	$V_1(\alpha, \beta)$	$V_2(\lambda, \mu)$

3.6 Epipolar line calculation

3.6.1 – CASE 1: two epipoles

a) theory

Let $M_1(x_1, y_1, z_1)$ be a point in P_1. E_1M_1 defines an epipolar line in P_1, and the corresponding E_2M_2 defines the conjugate epipolar line in P_2. The plane $(E_1, O_1, O_2, E_2, M_1, M_2)$ contains the translation vector t and E_1M_1. Its normal is $E_1M_1 \times t$. The normal of P_2 is O_2z. Hence the intersection of the two planes is given by the vector:

$$O_2z \times (E_1M_1 \times t) = (O_2z \cdot t)E_1M_1 - (O_2z \cdot E_1M_1)t$$

and E_2M_2 is collinear to this vector. Suppose that in O_1xyz, E_1M_1: $(x_1, y_1, 0)$. In terms of components in O_2xyz:

$$O_2z = \begin{pmatrix}0\\0\\1\end{pmatrix}, \quad t = \begin{pmatrix}\lambda\\\mu\\\nu\end{pmatrix}, \quad E_1M_1 = \begin{pmatrix}x_1'\\y_1'\\z_1'\end{pmatrix} = \begin{pmatrix}r_{11}x_1 + r_{12}y_1\\r_{21}x_1 + r_{22}y_1\\r_{31}x_1 + r_{32}y_1\end{pmatrix}$$

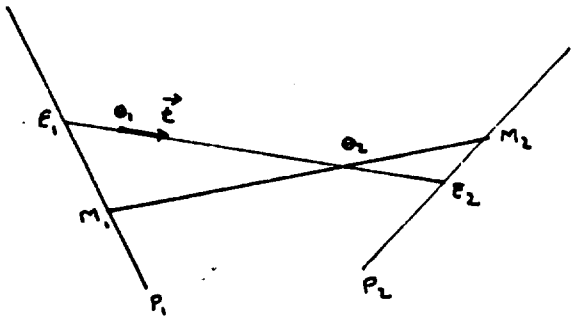
E_2M_2 is collinear to:

$$\begin{pmatrix}\nu x_1' - \lambda z_1'\\\nu y_1' - \mu z_1'\end{pmatrix} = \begin{pmatrix}x_1(\nu r_{11} - \lambda r_{31}) + y_1(\nu r_{12} - \lambda r_{32})\\x_1(\nu r_{21} - \mu r_{31}) + y_1(\nu r_{22} - \mu r_{32})\end{pmatrix}$$

If we let A be the matrix:

$$\begin{pmatrix}\nu r_{11} - \lambda r_{31} & \nu r_{12} - \lambda r_{32}\\\nu r_{21} - \mu r_{31} & \nu r_{22} - \mu r_{32}\end{pmatrix}$$

Then we can write $E_2M_2 = AE_1M_1$ where E_1M_1 is in base O_1xy and E_2M_2 is in base O_2xy.



Case 1
Figure 3-4

b) algorithm

Let N_1 be the number of epipolar lines that we want to determine. Each epipolar line is uniquely determined by the angle it makes with the x-axis. Let θ be this angle. Given k, the epipolar line number, $0 \leq k \leq n1 - 1$, how can we determine θ? If θ_0 and θ_1 are the lower and upper limits between which θ is allowed to vary, and $\theta_2 = \frac{(\theta_1 - \theta_0)}{N_1}$, then we will choose the middle of each interval: $\theta = \theta_0 + (k + .5)\theta_2$ But what are θ_0 and θ_1? We have to distinguish between three cases:

- The epipole is in the picture (very unlikely). Then θ can vary between 0 and 2π radians;
- The epipole is outside the image: there is a minimum and maximum angle under which the image is seen from this point. If we choose these angles in $[0, 2\pi]$, then most of the time every $\theta \in [\theta_0, \theta_1]$ will define a valid epipolar line in I_1;
- The exception from above is when the epipole is left of the image but on a same vertical level: then $[\theta_0, \theta_1]$ cannot be connected and still included in $[0, 2\pi]$. In this case we will choose the angles in $[\frac{-\pi}{2}, \frac{\pi}{2}]$.

Then L_1 will be defined by the point E_1 and the vector $V_1(cos\theta, sin\theta)$, and L_2 is defined by E_2 and $V_2 = AV_1$.

3.6.2 – CASE 2: one epipole E_1

a) theory

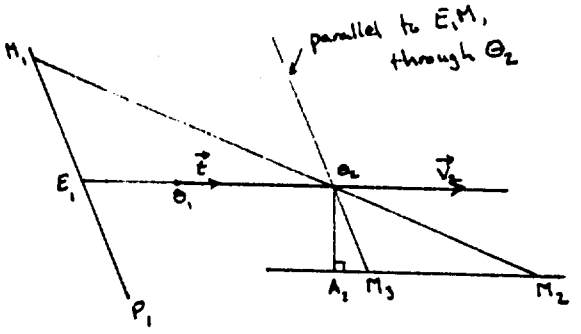
Given an epipolar line $\frac{L_1}{E_1M_1}$, we already know that the corresponding epipolar line L_2 in I_2 is collinear to the vector $t = V_2(\lambda, \mu)$. Hence we just need to find a point belonging to L_2. Clearly, L_2 is the intersection of I_2 with the plane (E_1, M_1, O_2). Thus any line contained in this plane will intersect I_2 at a point contained in L_2. In particular, consider the parallel to L_1 driven through O_2. It intersects I_2, thus L_2, at M_2 such that, in base O_2xyz:

$$E_1M_1 = \begin{pmatrix}r_{11}x_1 + r_{12}y_1\\r_{21}x_1 + r_{22}y_1\\r_{31}x_1 + r_{32}y_1\end{pmatrix}, \quad \text{hence} \quad O_2M_2 = \begin{pmatrix}f_2\frac{r_{11}x_1 + r_{12}y_1}{r_{31}x_1 + r_{32}y_1}\\f_2\frac{r_{21}x_1 + r_{22}y_1}{r_{31}x_1 + r_{32}y_1}\\f_2\end{pmatrix}$$

b) algorithm

In the same way as in case 1, we define $L_1(E_1, V_1)$ where $V_1(x_1 = cos\theta_1, x_2 = sin\theta_1)$. Then L_2 is defined by (M_2, V_2), where the coordinates of M_2 are

$$\left(f_2\frac{r_{11}x_1 + r_{12}y_1}{r_{31}x_1 + r_{32}y_1}, f_2\frac{r_{21}x_1 + r_{22}y_1}{r_{31}x_1 + r_{32}y_1}\right)$$



Case 2
Figure 3-5

3.6.3 – CASE 3: one epipole E_2

a) theory

Let an epipolar line in P_1 be defined by the translation vector $t = V_1$ and by a point M_1 we pick. In P_2, L_2 goes through E_2 and is collinear to a vector E_2M_2, intersection of P_2 with the plane (O_2, E_2, M_1). This plane is orthogonal to $O_2M_1 \times t$ and P_2 is orthogonal to O_2z. Hence the intersection is collinear to $O_2z \times (O_2M_1 \times t)$

$$\begin{aligned}\text{since } O_2M_1 &= O_2O_1 + O_1M_1\\&= kt + O_1M_1,\\O_2M_1 \times t &= O_1M_1 \times t,\\\text{and } O_2z \times (O_2M_1 \times t) &= (O_2z \cdot t)O_1M_1 - (O_2z \cdot O_1M_1)t.\end{aligned}$$

Suppose $O_1M_1 : (x_1, y_1, f_1)$ in O_1xyz. Then in O_2xyz:

$$O_2z = \begin{pmatrix}0\\0\\1\end{pmatrix}, \quad t = \begin{pmatrix}\lambda\\\mu\\\nu\end{pmatrix}, \quad O_1M_1 = \begin{pmatrix}x_1'\\y_1'\\z_1'\end{pmatrix} = \begin{pmatrix}r_{11}x_1 + r_{12}y_1 + r_{13}f_1\\r_{21}x_1 + r_{22}y_1 + r_{23}f_1\\r_{31}x_1 + r_{32}y_1 + r_{33}f_1\end{pmatrix}$$

E_2M_2 is thus collinear to:

$$\begin{pmatrix}\nu x_1' - \lambda z_1'\\\nu y_1' - \mu z_1'\end{pmatrix} = \begin{pmatrix}x_1(\nu r_{11} - \lambda r_{31}) + y_1(\nu r_{12} - \lambda r_{32}) + f_1(r_{13} - \lambda r_{33})\\x_1(\nu r_{21} - \mu r_{31}) + y_1(\nu r_{22} - \mu r_{32}) + f_1(\nu r_{23} - \mu r_{33})\end{pmatrix}$$

Hence, if we let A be the same matrix as in CASE 1 and OF3 be the offset matrix:

$$\begin{pmatrix}\nu r_{13} - \lambda r_{33}\\\nu r_{23} - \mu r_{33}\end{pmatrix}$$

Then we have: $E_2M_2 \cdot {}^tA\,O_1M_1 \wedge OP_2$

Where O_1M_1 is in base O_1xy and E_2M_2 is in base O_2xy

b) algorithm

Now, how do we pick M_1 in the first place? We want a set of N_1 equally spaced epipolar lines, and it appears convenient to pick points on the axes. If the epipolar lines are more horizontal, or the image stretched in height, then we will pick N_1 equally spaced points on the vertical axis, suitably located to cover the entire image. If the epipolar lines are more vertical or the image more stretched in width, then we pick them on the horizontal axis. Let L_x, L_y be the picture dimensions and (V_x, V_y) the epipolar direction:

If $V_xL_y < V_yL_x$, and $V_y < 0$, we pick
$$y_1 = (L_x|\tfrac{V_y}{V_x}| + L_y)(\tfrac{K+0.5}{N_1})$$

If $V_xL_y < V_yL_x$, and $V_y > 0$, we pick
$$y_1 = (L_x|\tfrac{V_y}{V_x}| + L_y)(\tfrac{K+0.5}{N_1}) - L_x\tfrac{V_y}{V_x}$$

If $V_yL_x < V_xL_y$, and $V_y < 0$, we pick
$$x_1 = (L_y|\tfrac{V_x}{V_y}| + L_x)(\tfrac{K+0.5}{N_1})$$

If $V_yL_x < V_xL_y$, and $V_y > 0$, we pick
$$x_1 = (L_y|\tfrac{V_x}{V_y}| + L_x)(\tfrac{K+0.5}{N_1}) - L_y\tfrac{V_x}{V_y}$$

Then we proceed as indicated above:
$L_1(M_1, V_1 = t)$ is matched with $L_2(E_2, V_2 = E_2M_2)$.

3.6.4 - CASE 4: no epipoles

a) theory

Let an epipolar line in P_1 be defined by the translation vector $t = V_1$ and by a point M_1 we pick. In P_2, L_2 goes through the image of M_1, that is the extremity of a vector collinear to O_1M_1 and whose third component is f_2. In base O_1xyz, $O_1M_1 : (x_1, y_1, f_1)$. In base O_2xyz:

$$O_1M_1 = \begin{pmatrix} r_{11}x_1 + r_{12}y_1 + r_{13}f_1 \\ r_{21}x_1 + r_{22}y_1 + r_{23}f_1 \\ r_{31}x_1 + r_{32}y_1 + r_{33}f_1 \end{pmatrix} \quad O_2M_2 = \begin{pmatrix} f_2\frac{r_{11}x_1 + r_{12}y_1 + r_{13}f_1}{r_{31}x_1 + r_{32}y_1 + r_{33}f_1} \\ f_2\frac{r_{21}x_1 + r_{22}y_1 + r_{23}f_1}{r_{31}x_1 + r_{32}y_1 + r_{33}f_1} \\ f_2 \end{pmatrix}$$

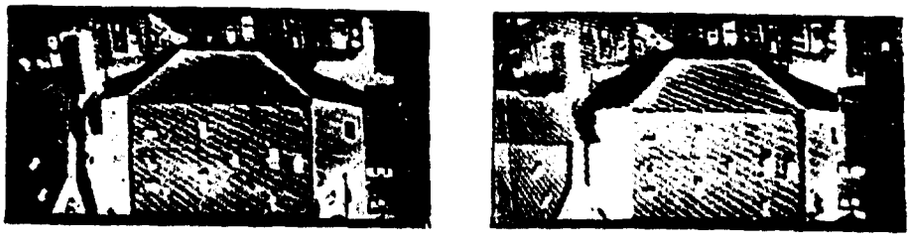
b) algorithm

We pick points E_1 in the same manner as in case 3. Then we calculate O_2M_2 as indicated above and we match $L_1(E_1, t)$ with $L_2(E_2, t)$.

3.7 Epipolar Registration and Transformation

Figure 3-6 shows a stereo pair of a building complex. Figure 3-7 has this pair superpositioned with a set of corresponding epipolar lines. Figure 3-8 shows the imagery transformed such that epipolar lines are horizontal in the image, and conjugate epipolar lines have the same row coordinate.

Sacramento Building Image Pair
Figure 3-6



Epipolar Lines in this Imagery
Figure 3-7

Transformed Imagery
Figure 3-8

4: Automated Stereo Mapping

4.1 Background

Results from our laboratory over the past few years [Quam 1971, Hannah 1974, Moravec 1980, Gennery 1980, Arnold 1980, Baker 1981, Arnold 1983], have demonstrated the possibilities of both area-based and feature-based stereo matching.

Area-based stereo matching uses windowing mechanisms to isolate parts of two images for cross-correlation. *Feature-based* stereo matching uses two-dimensional convolution operators (and perhaps grouping operators) to reduce an image to a depiction of its intensity boundaries, which can then be put into correspondence. Area-based cross-correlation techniques require distinctive texture within the area of correlation for successful operation. In general, it breaks track where there is no local correlation (zero signal, or where two images do not correspond, i.e. occlusions) or where the correlation is ambiguous (where the signal is repetitive).

Demands of mapping in cultural sites and in locales with surface discontinuity and ambiguous or non-existent texture make it essential that, if area-based analysis is to be done, it be done in conjunction with feature-based analysis. *Feature*-based analysis provides a solution to many of the problems of correlation. Principal among its advantages is that it operates on the most discriminable parts of an image: places that are distinctive in their intensity variation, and where localization is greatest. These are typically the boundaries between objects or between details on objects, or between objects and their backgrounds. The important point is that the features being put into correspondence for depth estimates are the boundaries of objects: area-based analysis is at its worst at object boundaries, yet determining boundaries can be said to be the most important part of mapping in 3-space.

The [Baker 1981] system is the only current system that mixes these two matching modalities. We have been working at applying this system to some cultural scenes. Before carrying out these analyses we wished to:

a) enhance the system with a capability to work with a better edge operator [Marimont 1982];

b) enable it to process images that are not graced with collinear epipolar geometry (*i.e.* most images);

c) introduce an additional correspondence measure - edge extent.

4.2 Epipolar Registration

To implement these enhancements required substantial redesign of the system, and redesign cycles with the Marimont process. Chosing useable data also presented difficulties, as the only imagery available was not of the correct geometry (see below). The two image pairs initially chosen (the Sacramento apartment complex and a section of some imagery of Moffett Field) proved, on closer examination, to require quite complex transformation, and could not be easily adjusted for epipolar processing.

In general, to bring imagery data into a properly transformed state could proceed in one of two ways:

- one could determine the transforms and then modify the imagery, producing an image pair having collinear epipolar geometry;

  or

- one could determine the transforms, and modify the output of an edge operator process that functions over the original imagery.

The latter is by far the superior approach, as it avoids resampling the image. This approach necessitates incorporating the transform computation into the stereo system, to follow edge finding and precede edge matching.

The second part of the stereo system's analysis is an intensity correlation process. This operates along epipolar lines as well, and clearly requires intensity information to be accessible along epipolar lines. One solution to this would be to take the original image pair and have the correlator rotate and change shape, size, and orientation as it moves around the image; this is an awkward and probably unnecessary complication. An alternative would be to access the transformed images, sampled as accurately as possible, and do the correlation in the rectangular space defined by collinear epipolar lines. The argument from edge accuracy indicated that transforming edges rather than resampling the image was the way to go; this argument from intensity correlation suggests that the resampled image can be useful.

Another implementation detail supported this use of both transformed edges and transformed imagery: it was found that the intensity information available from the Marimont process had too small a basis for useful correlation, and in fact, for transformed edges, had little relevance for the matching (it being measured not along epipolar lines, but normal to the edge direction). The transformed image had to be referenced again by the system to obtain more significant intensity estimates oriented along epipolar lines, and working with the image in epipolar space facilitated this.

The philosophy of the stereo matching process here had been to use edge analysis for, among other things, its higher accuracy, and to use intensity analysis for the continuity it provides. To be consistent with this, we wanted to have the highest possible accuracy for edges in epipolar space, and if sacrifice be needed for simplicity, to do it where it least degraded the analysis - in the intensity correlation. It is clear that transformed edges give higher accuracy than edges from transformed images (detectability might not change much, but localization is significantly reduced); and important simplifications could be obtained for little loss by doing the intensity correlation over the resampled image pair. This meant changes in our plans for the registration system: it had to produce not just transform information, but transformed images as well. Both forms are made available as output from the registration program described in section 3, and the enhanced Baker system uses them both.

4.3 The Marimont Edge Operator

The Marimont edge operator has greater detection and reliability than the original Baker edge operator, and similar localization; earlier examples of its processing convinced us that its output would improve the quality of our stereo reconstruction. Its ability to track along zero signal areas in following zero-crossing edges leads to more coherent image descriptions. [Marimont 1982] provides details of the operator's functioning. Roughly, it works by convolving an $m \times m$ lateral inhibition function of $n \times n$ central window with an image. Zero crossings in this resultant image then indicate edges, and the edge position is determined by interpolating over the lateral inhibition surface.

A few unanticipated problems became apparent once work with the edges was begun. One point, noted above, was that the intensity information stored at an edge (its left and right boundary values) had quite small support (a single pixel). This is in contrast with the original operator which interpolated for these values in an area 3 pixels wide and removed one pixel from the determined edge position. Another problem was that the edge connectivity produced by the Marimont system can be misleading. Intensity significance was improved by sampling along epipolar lines in the transformed images. The connectivity problem has not been looked at yet. Good connectivity is inherently difficult to achieve with zero crossing operators. Refinements to the process are being considered.

4.4 Edge Extent

The introduction of edge extent as a parameter in the dynamic programming solution was an obvious fallout from using the Marimont edges. Edges are output by that process as strings, with 2-connectedness. The maximum and minimum of some string, in transform space, is a measure of its (epipolar) extent. Prior to the use of this information the only way that global continuity entered the analysis was through a consistency enforcement relaxation process which ensured that edges connected in one view were interpreted as continuous in 3-space; all matching measures were quite local. With the modified approach, the correspondence measure is a function of (among other, more statistically based parameters) the ratio of edge extents. In particular, the likelihood of edge element a in the left image matching edge element b in the right image depends on the product of the ratios of the two upper extents (up from the edge elements) and the two lower extents (down from the two edge elements).

4.5 Testing on Imagery

When image testing began with all of the above accomplished, another problem became apparent: the stereo system, bound into a machine architecture with a maximum of 256K words of memory, and always tightly wedged anyway, had grown with these changes to the point that only small portions of images could be worked on at once. This came to exist a windowing mechanism within the edge finding/loading and stereo matching processes.

Our testing has been progressing on several sets of imagery: a synthetic image pair from Control Data Corporation, an aerial scene from the Engineering Topographic Laboratory, and a building scene of Sacramento. We will report on the results of these analyses at a later date.

5: Summary

A principal research interest of our group is in developing a rule-based advanced automated stereo mapping system to function within ACRONYM [Brooks 1981]. Current mapping techniques ignore much of the information available from inference on single views of a scene. This information can be useful for three-dimensional surface interpretation, and also provides extra parameters for stereo matching (i.e. surface orientation, occlusion cues). Our research effort is directed at establishing such monocular inference rules in a rule-base for stereo mapping.

In deriving these rules, we perform analysis of both hand extracted and automatically produced edge descriptions. A facility has been developed for this manual edge extraction from hardcopy imagery. We have studied rule synthesis for several cases, including that of orthogonal trihedral vertices - features that dominate cultural scenes. This research is very promising, and has shown the utility of the rule-based approach to surface inference from monocular information.

Camera solving provides powerful constraint on the correspondence problem in stereo matching. We have developed a facility for interactively registering images, determining the parameters for transforming them (or their edge descriptions) into collinear epipolar space, and performing the actual image transformation. This determination is crucial to a mapping process. Incorporating an automated module to provide data for the camera solving is a very important next step.

We have experimented with an existing stereo mapping process, enhancing its flexibility with respect to image format and with respect to edge operator format, and have been preparing example outputs of its processing on new imagery. Our intent with this effort has been to show the capabilities of a local matching process and to assess its applicability to the planned rule-based system.

Acknowledgements

This work was supported by Syracuse University Subcontract PO # 20205-135 (Air Force F30602-81-C-0169), NASA NAG 5261, and ARPA Contract N0039-82-C-0250.

6: References

[Arnold 1980] Arnold, R.D., and T.O. Binford, "Geometric Constraints in Stereo Vision," *Soc. Photo-Optical Instr. Engineers*, Vol. 238, Image Processing for Missile Guidance, 1980, 281-202.

[Arnold 1983] Arnold, R. David, "Automated Stereo Perception," Department of Computer Science, Stanford University, Ph.D. thesis, 1983.

[Baker 1981] Baker, H. Harlyn, "Depth from Edge and Intensity Based Stereo," University of Illinois, Ph.D. thesis, September 1981.

[Binford 1981] Binford, Thomas O., "Inferring Surfaces from Images," *Artificial Intelligence*, Vol. 17(1981), August 1981, 205-244.

[Brooks 1981] Brooks, Rodney A., ' Symbolic Reasoning Among 3-D Models and 2-D Images," Stanford Artificial Intelligence Laboratory, AIM 343, June 1981.

[Duda 1973] Duda, R.O. and P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[Gennery 1980] Gennery, Donald B., "Modelling the Environment of an Exploring Vehicle by Means of Stereo Vision," Ph.D. thesis, Stanford Artificial Intelligence Laboratory, AIM 339, June 1980.

[Hallert 1960] Hallert, Bertil, *"Photogrammetry, Basic Principles and General Survey,"* McGraw-Hill Book Company Inc., 1960.

[Hannah 1974] Hannah, Marsha Jo, "Computer Matching of Areas in Stereo Images," Ph.D. thesis, Stanford Artificial Intelligence Laboratory, AIM-239, July 1974.

[Liebes 1981] Liebes Jr., S., "Geometric Constraints for Interpreting Images of Common Structural Elements: Orthogonal Trihedral Vertices," *Proceedings of the DARPA Image Understanding Workshop*, 1981.

[Lowe 1982] Lowe, David G., "Segmentation and Aggregation," *Proceedings of the DARPA Image Understanding Workshop*, Stanford University, September 1982.

[Marimont 1982] Marimont, David H., "Segmentation in ACRONYM," *Proceedings of the DARPA Image Understanding Workshop*, Stanford University, September 1982.

[Moravec 1980] Moravec, Hans P., "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," Stanford Artificial Intelligence Laboratory, AIM-340, Ph.D. thesis, September 1980.

[Perkins 1968] Perkins, "Cubic Corners," in Quarterly Progress Report, MIT Electronics Laboratory, April 15,1968

[Quam 1971] Quam, Lynn H., "Computer Comparison of Pictures," *Stanford Artificial Intelligence Laboratory, AIM-144*, Ph.D. thesis, 1971.